SHARED LINGUISTIC RESOURCES FOR HUMAN LANGUAGE TECHNOLOGY IN THE MEETING DOMAIN

Stephanie Strassel, Meghan Glenn

Linguistic Data Consortium

ABSTRACT

This paper describes efforts by the University of Pennsylvania's Linguistic Data Consortium to create shared linguistic resources in support of the 2004 NIST Meeting Recognition Evaluation. LDC created training transcripts for thirteen hours of meetings from NIST's Pilot Meeting Corpus, which represents a wide variety of subjects, scenarios and recording conditions. The data was transcribed using a Quick Transcription methodology, an approach that sacrifices some quality for maximum efficiency. In addition to training data, LDC also produced evaluation transcripts for the ninety minutes multi-site evaluation corpus. In contrast to the training data, the evaluation transcripts were produced using a Careful Transcription specification where every file receives at least three separate passes, each focusing on a different annotation task. The paper describes both the quick and careful transcription approaches, assessing the pros and cons of each strategy. Finally, the paper touches on LDC's meeting data collection activities and the resulting contributions of meeting speech to the 2004 evaluation corpus.

1. INTRODUCTION

All human language technology requires large volumes of data for system training and development as well as stable benchmark data to measure ongoing progress. The Linguistic Data Consortium (LDC) was founded in 1992 at the University of Pennsylvania, with seed money from DARPA, specifically to address the need for shared language resources to support research, education and technology development. As part of NIST's 2004 Meeting Recognition Evaluation, LDC produced reference transcripts of training and evaluation data to support automatic speech-to-text transcription and speaker segmentation in the meeting domain; LDC also contributed meeting sessions to the pool of multi-site evaluation data.

2. TRANSCRIPTION METHODOLOGY

The cost of producing careful manual transcripts in sufficient quantity to provide data for system training and development can be quite high. Careful transcription rates approach twenty times real time per channel, so that it can requires twenty or more hours of annotator effort to carefully transcribe one hour of single-channel speech. In the meeting domain where single sessions can easily include a half dozen speakers, the cost of carefully transcribing large volumes of data can be prohibitively expensive. In order to reduce both costs and turnaround time in producing reference transcripts for the NIST Pilot Meeting Corpus, LDC employed two different strategies to create training data versus evaluation data.

2.1 Careful Transcription: Evaluation Data

For purposes of evaluating speech technology systems, system output must be compared with high-quality manually-created verbatim transcripts. LDC used a careful transcription approach in creating the evaluation reference transcripts for the 2004 Meeting Evaluation. (A quarter of the evaluation data had already been transcribed by the contributing site, ICSI; LDC staff modified and checked the existing transcripts to bring them into line with our careful transcription approach.) The careful transcription effort involves multiple passes over the data. Annotators first manually segment speaker turns and (for broadcast data) story boundaries, as well as indicating smaller breakpoints within the audio stream that correspond to breath or pause groups. After accurate segment boundaries are in place, annotators create a verbatim transcript by listening to each segment in turn. A second pass checks the accuracy of the segment boundaries and transcript itself, revisits difficult sections, and adds information like speaker identity, background noise conditions, plus special markup for mispronounced words, proper names, acronyms, partial words and the like. Further scans over the data identify common errors, conduct spelling and syntax checks, and standardize the spelling of personal, organization and other names across the transcripts.

2.2 Quick Transcription: Training Data

The training data comprised thirteen hours of multi-channel recordings from NIST's Pilot Meeting Corpus, with a wide variety of subjects, scenarios and recording conditions. This data was transcribed using a Quick Transcription (QTR) methodology, originally developed as part of LDC telephone transcription efforts [1]. The goal of QTR is to provide a content-accurate transcript, sacrificing some quality and extra markup to produce the data as quickly as possible. Rather than executing three to four separate passes over the data, annotators create a (nearly) verbatim transcript in a single pass. Automatic post-processing targets spell checking, syntax checking and scans for common errors. Team leaders monitor annotator progress and speed to ensure that transcripts are produced within the targeted timeframe. The resulting quick transcription quality is naturally lower than that produced by the careful transcription methodology. Speeding up the process inevitably results in missed or mis-transcribed speech; this is particularly true for difficult sections of the transcript, including disfluent or overlapping speech sections. However, the advantage of this approach is undeniable. Annotators work, on average, ten times faster using this approach than they are able to work within the careful transcription methodology.

To expedite the transcription process and to accommodate a severely compressed timeline, LDC made the decision to modify our typical QTR approach and instead outsource the initial transcription of the training data to a professional transcription agency, who could complete transcription of the full data set in three days. The external agency executed a single pass over each recording, relying on only one channel of audio (a mix of the individual speakers' head mounted microphones) to create a basic transcript of each session. In addition, the external agency was asked to assign speaker IDs to each turn. The agency was provided with samples from the individual head-mounted mic recordings for each speaker to aid transcribers in identifying each speaker's voice and assigning the correct ID. After receiving the data back from the agency, LDC annotators would then manually time-align the base transcripts to the audio, review the transcript content, add minimal markup and verify speaker IDs. These seemingly straightforward measures in fact presented numerous unexpected challenges.

Typically, transcription begins with segmentation; that is, virtually chopping the audio into smaller units. This makes later transcription easier by presenting the annotator with a series of small units to transcribe, rather than a large undifferentiated stream of speech. However, because the training data had been pre-transcribed without time alignment, LDC transcribers needed to align the existing text with the corresponding audio. We experimented with three different methods for doing this.

The first strategy adopted was a target segment approach. Using a single-channel recording (the head mounted mic mix), senior annotators segmented the speech into individual speaker segments. The segments were necessarily approximate in some cases, given the overwhelming presence of overlapping speech. The existing transcripts were then matched up with the resulting segment boundaries, speaker by speaker and turn by turn. A second pass over the data required transcribers to listen to the individual speaker recordings and refine the timestamps and transcripts as needed.

A modified version of this basic approach was also attempted, starting with the individual speaker recordings rather than the multi-speaker mix. However, we quickly came to realize that this approach was not effective. The primary problem was in the serious inadequacies of the professionally-created transcripts. The quality of the transcripts on the whole was quite poor, with entire speakers left untranscribed or undertranscribed, so that LDC transcribers were required to re-transcribe large sections of each session. In addition, more than half of the speaker IDs assigned by the transcription agency were incorrect, due to overlapping speech regions where the external agency had transposed, merged or entirely missed individual speakers.

Because LDC staff were required to spend considerably more time than had been budgeted in correcting transcripts and speaker IDs, we needed to expedite the time alignment as much as possible. To achieve this, we relied on an approach that had been applied previously with much success. In the QTR methodology, annotators do not manually segment the speech file into turns or smaller units. Instead, an automatic process developed at LDC pre-segments a speech file into high-accuracy turn boundaries. This is optionally followed by human verification of the timestamps, a procedure that requires approximately 1.5 times real time per channel to complete. The same automatic methods were applied to this problem. We applied our AutoSegmenter to the individual speaker recordings and created high-accuracy segment boundaries for each speaker. Transcribers then reviewed the existing transcripts, matching up as much of the text to the speech as possible and creating new transcripts for the un(der)transcribed sections. Senior transcribers then conducted two quality checks on the resulting time-aligned transcripts. One pass used the single speaker recordings to verify timestamp accuracy and transcript completeness. After the individual speaker transcripts were complete, another automated process merged the transcripts together, and the team leader conducted a final QC pass to check spelling, look for common transcription errors, standardize names, verify timestamps and ensure that no speech had gone unsegmented or untranscribed.

3. UNIQUE CHALLENGES OF MEETING SPEECH TRANSCRIPTION

The meeting domain presents some added challenges to more traditional speech domains, not only to researchers, but to corpus creators. The fundamental challenge in transcribing the data is simply the added volume resulting from not one or two but a half dozen or more speakers. While a typical thirty-minute telephone conversation may require twenty hours or more to transcribe carefully (30 minutes, 2 speakers, 20 times real time per channel), a thirty-minute meeting with six participants may require more than 60 hours to produce a transcript of the same quality. Methods like Quick Transcription can cut these times considerably, but the volume of effort required is still substantial.

Even when working with the individual speaker recordings, overlapping speech is a serious challenge. Transcribers must focus their attention on a single speaker's voice, while simultaneously considering the context of the larger conversation to understand what is being said. This was particularly true in the NIST Pilot Meeting Corpus, where LDC transcribers encountered highly technical and project-specific vocabulary and acronyms as well as personal names and nicknames. Another aspect of meetings that initially may not seem like a hurdle but in fact proved to be both difficult and frustrating for transcribers is that some meeting participants do not engage in the conversation. Annotators may need to listen to long stretches of silence, coughing and sighing before they hear any speaker utterances at all. Some of the more amusing meetings presented their own challenges for transcribers. For example, one of the recordings was of several participants playing a board game. This required annotators to listen to long segments of game pieces clicking across a board, of money being counted, of whispered verbal taunts, and of dice rolling (followed by the inevitable “one, two, three, four, five” of spaces being counted). Another game forum introduced the transcription crew to a highly specific storytelling game with its own vocabulary words, which presented difficulties even after transcribers memorized the game manual.

The nature of meeting speech transcription requires nearly constant jumping back and forth from a single speaker to the multi-speaker view of the data, which presents a challenge not only for the transcribers, but for the transcription tools they use. Most current tools assume either one-channel, multi-speaker data (as in the broadcast news domain), or two channel, single speaker per channel data (as in telephone speech). The ideal meeting data transcription tool would merge features of each, allowing users to easily move back and forth between the multi- and single-speaker views, turning individual channels on and off as required to customize their interaction with the data. While LDC is currently developing such a tool, called XTrans [2], until it is in place transcribers must make do with non-optimal solutions, which adds time and effort to a task that is already substantial.

4. MEETING DATA COLLECTION AT LDC

In addition to creating training and evaluation transcripts to support the NIST Meeting Evaluation, LDC also contributed two sessions to the evaluation data pool. These sessions were collected as part of an LDC initiative in support of the ROAR (Reliable Omnipresent Automatic Recognition) Project. LDC recruited English speaking participants to participate in two kinds of recording sessions. GroupMeet was targeted at groups of speakers who were willing to hold an already-planned meeting in LDC's meeting recording facility. GroupTalk was constructed as a facilitated discussion, where one initially-contacted subject selects the other participants. The GroupTalk approach was designed to minimize the effect of the interviewer and maximize comfort and the flow of conversation. The interview topics were also designed to facilitate the flow of speech. A sample question in that domain is "Did you ever get blamed for something you didn't do?" The research builds on previous LDC data collection projects, such as CallFriend and CallHome, which have been widely used in speech recognition research. The design of ROAR also builds on the pioneering work of William Labov and others at the University of Pennsylvania in the area of sociolinguistic methodology, including the development of the standard sociolinguistic interview. The data resulting from ROAR was designed to be of the high quality required by the speech recognition community, while the interview format itself was designed to elicit the natural, spontaneous speech used by sociolinguists. The interview sessions were recorded in multi-track format on computer disk as well as on digital tape for archival purposes. The microphones used were wireless lavalier, headset microphones, freestanding and wall-mounted condenser microphones. The goal of the ROAR effort was to obtain naturalistic, conversational speech under the best possible conditions for audio recording. The two sessions selected for contribution to the Meeting Evaluation data pool were chosen for the variety of microphones available and the high quality of the resulting speech data.

5. DATA DISTRIBUTION AND PUBLICATION

In addition to creating data, LDC's primary mission is to distribute resources to the researchers who need them. In support of the NIST Meeting Evaluation, LDC expedited general publication of the ICSI Meeting Corpus [3, 4] to ensure its availability to evaluation participants.

Because general release publications typically require a month or more to process, validate and replicate for distribution, LDC has developed a new data distribution method known as eCorpora specifically to allow for expedited delivery of data to a limited number of research sites participating in common task evaluations. LDC used this eCorpus method to distribute the NIST Pilot Meeting Corpus to evaluation participants. The ISL Meeting Corpus, also used as training data, was also distributed by LDC using this method. Upon the conclusion of the formal task evaluation, pending negotiations with research sponsors and program coordinators, LDC publishes eCorpora more broadly to permit access to these valuable resources to all communities working in linguistic education, research, and technology development. Both the ISL and the NIST Pilot Meeting Corpus are scheduled as general LDC publications in 2004.

6. REFERENCES

[1] S. Strassel, C. Cieri, K. Walker and D. Miller, “Shared Resources for Robust Speech-to-Text Technology,” Proceedings of Eurospeech 2003, Geneva, Switzerland, September 2003

[2] K. Maeda, S. Strassel, "Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium," to appear in the Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 2004.

[3] A. Janin, et. al. ICSI Meeting Corpus Speech. Linguistic Data Consortium Catalog Number LDC2004S02 (ISBN 1-58563-285-6). February, 2004.

[4] A. Janin, et. al. ICSI Meeting Corpus Transcripts. Linguistic Data Consortium Catalog Number LDC2004T04 (ISBN 1-58563-286-4). February, 2004.